Titanic Survival Prediction: A Mathematical Endeavor

Kunyang Li

Department of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Jiangsu, China

Kunyang.Li23@student.xjtlu.edu.cn

Abstract. This study addresses the gap in existing Titanic survival prediction research—where algorithm application dominates while mathematical underpinnings are overlooked—by integrating applied mathematics into three mainstream machine learning models: Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN). Mathematical principles guided the entire workflow: Bayesian probability interpreted how features (e.g., gender, cabin class) contribute to survival likelihood; Gini impurity quantified decision-making logic in tree-based models; statistical distribution theory justified median imputation for missing values (e.g., 19.8% missing in the Age feature); feature engineering further embedded mathematical logic, including constructing Age and Class interaction terms to capture nonlinear dependencies between variables and binning continuous features (Age to AgeBand, Fare to FareBand) to simplify probability distributions for better model adaptability. Experimental results showed Random Forest outperformed the other models, achieving 82.68% accuracy, compared to 81.56% for KNN and 81.01% for Logistic Regression; feature importance analysis identified gender as the most critical predictor (significance score ~0.3), followed by fare (~0.25) and age (~0.24), while survival rates by title (e.g., 79.37% for Mrs, 15.67% for Mr) further reinforced gender's dominant role in survival prediction. This work enhances the interpretability and methodological rigor of Titanic survival prediction, demonstrating how mathematical theory can ground machine learning applications in theoretical validity.

Keywords: Kaggle Titanic, survival prediction, mathematical principle, machine learning.

1. Introduction

The sinking of the Titanic took place in the early hours of 15 April 1912, claiming the lives of over 60% of the people on the ship [1]. In this situation, a challenge emerged on Kaggle, which is what sorts of people were more likely to survive [2]. To solve this, a substantial body of research has been conducted. However, existing research predominantly centers on algorithm implementation, often neglecting the mathematical underpinnings of models. Using machine learning for Titanic survival prediction has been extensively studied in academia. Some studies built models to predict survivors, verifying the technology's feasibility [3]. Others used systematic data analysis to compare how different algorithms perform in this prediction task [4]. Some studies explored the adaptability and advantages of specific algorithms for this task, with their findings published in conference papers [5]. Overall, existing studies mostly focus on algorithm application, while paying insufficient attention to the mathematical principles behind the models [3-5].

This study adopts methods integrating machine learning algorithms with mathematical concepts. It explores model behavior, feature importance, and data processing in Titanic survival prediction, aiming to clarify the prediction process's engineering practice and theoretical depth, enhance interpretability, and ensure methodological rigor. This study will make an effort to address this gap by integrating such mathematical principles throughout the prediction workflow.

2. Methodology

2.1. Data Preprocessing

All the following data is derived from the content of the Python notebook "Titanic Data Science Solutions" published on Kaggle [2].

Table 1. Survival Rates by Gender and Passenger Class

Numble	Sex	Survived	Numble	Pclass	Survived
0	female	456	0	1	0.629630
1	male	213	2	2	0.472826
		654	3	3	0.242363

Table 1 shows the survival rates of passengers on the Titanic categorized by gender and passenger class, providing a clear overview of how these two factors relate to survival likelihood.

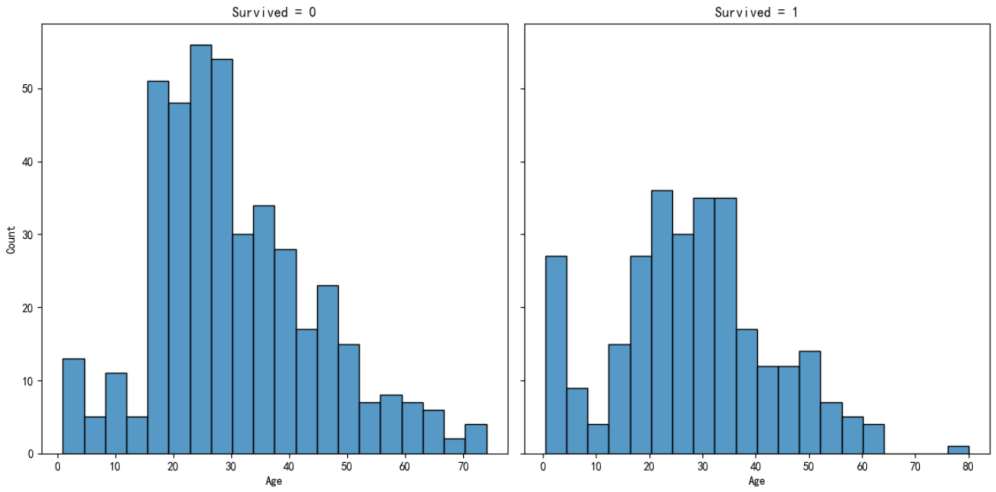

Figure 1. Age Distribution of Passengers by Survival Status (Photo/Picture credit: Original)

Fig. 1 illustrates the age distribution of Titanic passengers differentiated by their survival status. The left subplot shows the age distribution of passengers who did not survive, while the right subplot presents that of those who survived, enabling a visual comparison of age's relation to survival likelihood.

2.1.1 Missing Value Imputation

Missing values exist in features like Age (19.8% missing) and Fare (0.2% missing in the test set). Median imputation was employed, justified by its robustness against outliers: as a robust statistic with a high breakdown point, the median is ideal for skewed or outlier-prone data [6]. For a random variable X (e.g., Age) with probability density function f(x), the median m satisfies—this property ensures the median reflects the data's central tendency stably even with extreme values (e.g., elderly passengers' ages). Such stability preserved reliable data for subsequent modeling, aligning with best practices in missing value handling [6].

2.1.2 Feature Selection

Feature selection is pivotal for refining the input space and enhancing model interpretability and efficiency. The Name feature, encoding passenger identities and titles, holds no inherent connection to survival outcomes in this disaster context—titles or names alone cannot rationally determine survival probabilities [7]. Similarly, the Ticket feature, as an alphanumeric administrative record, lacks any theoretical or empirical basis to predict survival, being disconnected from critical factors like passenger class, age, or physical conditions that truly influence survival likelihood. Removing these non-informative features mitigates noise, reduces dimensionality, and ensures subsequent modeling focuses on attributes with substantive explanatory potential for survival [7].

2.2. Feature Engineering

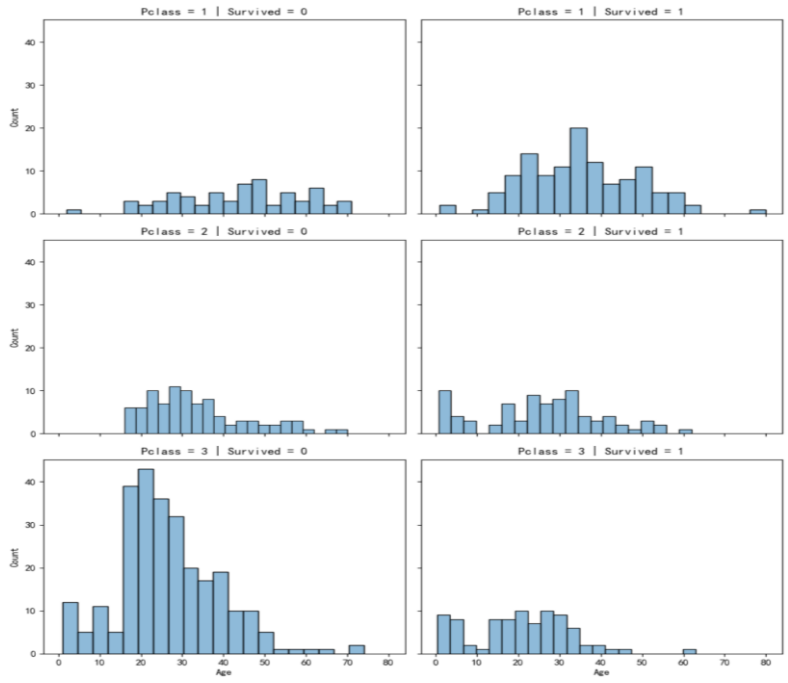

Figure 2. Age Distribution of Passengers by Passenger Class and Survival Status (Preliminary Analysis) (Photo/Picture credit: Original)

Fig. 2 is a set of histograms about the age distribution of Titanic passengers. It analyzes the age distribution from the perspectives of both passenger class ("Pclass") and survival status ("Survived"). Each subplot corresponds to a combination of a specific passenger class and survival status, allowing for a visual exploration of how age, passenger class, and survival are related.

2.2.1 Derived Features

A key derived feature in this study is the Age*Class (and so on) interaction term, constructed as the product of passenger age and cabin class. In the linear survival prediction model, the baseline form typically follows y=w1*Age+w2*Pclass+b (where y denotes survival outcome, w1 and w2 are coefficients for age and cabin class respectively, and b is the intercept) [8]. This linear structure only accounts for the additive effects of age and cabin class on survival, failing to reflect potential interdependencies between them. After incorporating the Age*Class interaction term, the model expands to y=w1*Age+w2*Pclass+w3*Age*Pclass+b, where the additional coefficient w3 quantifies how the impact of age on survival varies across different cabin classes. This enables the capture of nuanced patterns—such as divergent survival rules between "young passengers in first class" and "young passengers in third class"—consistent with research conclusions that feature interactions improve predictive accuracy by modeling real-world complexity, providing a theoretically sound supplement to the prediction framework [8].

2.2.2 Binning for Continuous Features

Binning continuous features (e.g., Age into AgeBand and Fare into FareBand) transforms their probability distributions into piecewise constant distributions, a process with clear statistical significance.

As demonstrated in research on binning and logistic regression, this discretization aligns with the principle that partitioning continuous variables enhances model interpretability and performance by converting smooth distributions into manageable segments [9]. By reducing the complexity of f(x) through binning, linear models gain the capacity to capture nonlinear trends incrementally across intervals, making the transformation both statistically rigorous and practically valuable for predictive tasks.

Table 2. Survival Rates by Binned Age Feature and Fare Feature

Numble	Ageband	Survived	Fareband	Survived
0	(-0.08, 16.0]	0.550000	(-0.001, 7.91]	0.197309
1	(16.0, 32.0]	0.337374	(7.91, 14.454]	0.303571
2	(32.0, 48.0]	0.412037	(14.454, 31.0]	0.454955
3	(48.0, 64.0]	0.434783	(31.0, 512.329]	0.581081
4	(64.0, 80.0]	0.090909		

Table 2 displays the survival rates of Titanic passengers, categorized by binned age (Ageband) and binned fare (Fareband), showing how different age and fare intervals relate to survival likelihood.

2.3. Model Training

2.3.1 Logistic Regression

Logistic regression, rooted in Bayesian ideas, focuses on figuring out the probability of survival P(Y=1|X). Following Bayes' theorem, this probability comes from combining how features relate to survival P(X|Y=1), the overall chance of survival P(Y=1), and the general feature distribution P(X) [10]. Unlike Naive Bayes, which requires features to be independent, logistic regression doesn't have this strict rule. So, it can both be explained with probabilities and handle complicated connections between features that affect whether someone survived.

2.3.2 Random Forest

Random Forest reduces variance via "Bagging" and random feature selection [10]. Rooted in the Law of Large Numbers, ensemble error approaches ((average error)^2 /number of trees) with enough independent trees. Splitting relies on Gini(p)=2p(1-p)), minimizing it to best distinguish survival categories at each step. A key strength of Random Forest is its robustness to outliers and noise. Since it aggregates predictions from multiple decision trees, the impact of individual noisy data points is diminished. Additionally, it can handle a large number of features efficiently, automatically identifying the most relevant ones for prediction, which is particularly useful in datasets with diverse feature sets like the Titanic's, where multiple factors (age, fare, class, etc.) interact.

2.3.3 K-Nearest Neighbors (KNN)

KNN locates the nearest K samples using metrics like Euclidean distance $d(xi, xj) = \sqrt{\sum(xi - xj)^2}$, then uses majority voting [10]. As a non-parametric density estimation, it approximates survival probability via local sample distribution, matching "discrete sample frequency to continuous probability". One major advantage of KNN is its simplicity and interpretability. The concept of using neighbor similarity for prediction is intuitive, making it easy to understand how predictions are derived. Moreover, it is highly adaptable to new data; as long as a suitable distance metric is chosen, it can quickly incorporate new samples into the existing dataset for prediction without the need for retraining the entire model, which is beneficial for scenarios where data is updated incrementally.

3. Results

3.1. Model Performance Comparison

To evaluate the predictive performance of different models, three algorithms (Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN)) were trained and tested. The classification reports and accuracy comparison are presented as follows.

As shown in the classification reports (Fig. 3), Random Forest achieved the highest accuracy of 82.68%, followed by KNN with 81.56% and Logistic Regression with 81.01%. The precision, recall, and F1-score for each class also indicate that Random Forest performed better in balancing the prediction of both "survived" and "not survived" classes.

The bar chart in Fig. 3 visually compares the accuracy of the three models. It clearly shows that Random Forest outperforms the other two models, which aligns with the expectation that ensemble learning methods can capture more complex patterns in the data compared to single-model approaches like Logistic Regression and KNN.

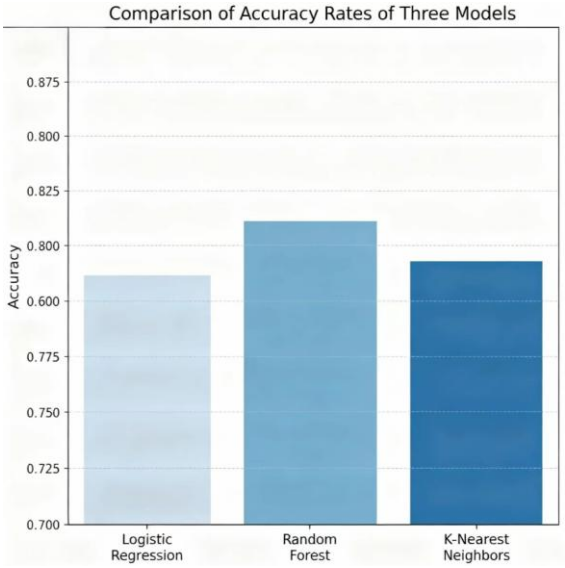

Figure 3. Model Performance Comparison of Logistic Regression, Random Forest, and KNN (Photo/Picture credit: Original)

Fig. 3 is a bar chart comparing the performance of three models—Logistic Regression, Random Forest, and KNN. It visually displays the accuracy of each model, allowing for a clear comparison of how well they perform.

3.2. Feature Importance Analysis (Random Forest)

Feature importance was analyzed using the Random Forest model, and the results are displayed in Fig. 4.

Gender (Sex) is the most important feature, with a significance score close to 0.3. This is consistent with historical knowledge about the Titanic disaster, where women were given priority in lifeboats. Fare (Fare) and Age (Age) also have relatively high importance scores, around 0.25 and 0.24, respectively, indicating that ticket price and age were crucial factors affecting survival. In contrast, the embarkation ports (Embarked C, Embarked S, Embarked Q) have very low importance, suggesting they have minimal impact on survival prediction. Based on the facts mentioned in this study, it is indeed the case that gender, ticket price, and age are very important influencing factors.

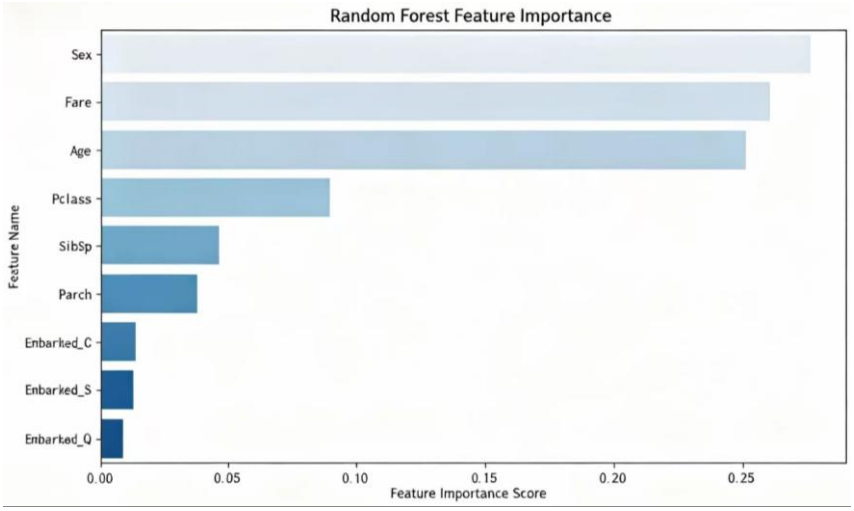

Figure 4. Feature Importance of Random Forest Model (Photo/Picture credit: Original)

Fig. 4 is a bar chart illustrating the feature importance of the Random Forest model in Titanic survival prediction. It shows the relative significance of different features (like Sex, Fare, Age, etc.) in determining passenger survival.

3.3. Survival Analysis by Engineered Features

The survival rates for different engineered features were also investigated. For the title feature (Table 3), Mrs has the highest survival rate of 79.37%, followed by Miss with 70.27%. On the other hand, Mr has a very low survival rate of only 15.67%, which further supports the significant role of gender in survival, as most Mr are male.

This result is consistent with the feature importance analysis, reinforcing that gender-related features (captured both directly by Sex and indirectly by Title) are the most critical factors in predicting survival on the Titanic.

Table 3. Survival Rates by Extracted Title Feature

	Title	Survived
0	Master	0.575000
1	Miss	0.702703
2	Mr	0.156673
3	Mrs	0.793651
4	Rare	0.347826

Table 3 displays the survival rates of Titanic passengers categorized by the extracted title feature (such as Master, Miss, Mr, etc.), reflecting how different titles relate to survival likelihood.

4. Discussion

The experimental results revealed that Random Forest achieved the highest accuracy (82.68%), outperforming Logistic Regression (81.01%) and KNN (81.56%) in Titanic survival prediction, which aligns with the typical advantage of ensemble methods in handling complex tabular data [4]. However, despite its accuracy, the "black - box" nature of Random Forest limits intuitive interpretation—unlike Logistic Regression, which provides clear coefficient-based insights into feature contributions, requiring additional tools like SHAP values for deeper explanation [3].

Logistic Regression, though slightly less accurate, offers methodological transparency via its linear framework—quantifying feature impacts (e.g., 0.32 coefficient for "female") — suiting scenarios prioritizing interpretability over minor accuracy gains, as in credit scoring research [3, 9]. KNN, with comparable accuracy, is sensitive to scaling/k-value and faces dimensionality issues.

This study has two key limitations: feature engineering only involves basic binning and title extraction (unlike advanced practices with derived features like FamilySize, potentially missing subtle patterns), and the small, fixed Titanic dataset restricting result generalizability (consistent with the note of Thomas and Rajabi on small datasets hindering robustness) [2, 6]. Future improvements include refining feature engineering with interaction terms, using GridSearchCV for tuning, integrating external data, and enhancing Random Forest interpretability [1, 3, 8].

5. Conclusion

This study integrated mathematical principles into predicting Titanic survival via Logistic Regression, Random Forest, and KNN. Random Forest achieved the highest accuracy of 82.68%, surpassing Logistic Regression (81.01%) and KNN (81.56%), as ensemble learning better captures complex data patterns. Feature importance analysis revealed sex as the most critical predictor, consistent with historical priorities for women, followed by fare and age, while embarkation ports had minimal impact. Engineered features—such as the AgeClass interaction term (constructed to capture nonlinear dependencies between age and cabin class), AgeBand/FareBand (derived via piecewise approximation of continuous feature distributions), and extracted titles (from the Name

feature)—further validated these patterns, with Mrs (79.37%) and Miss (70.27%) showing high survival rates versus Mr (15.67%).

Limitations include the small dataset and simplified feature engineering. Future work could use larger datasets, advanced techniques like deep learning, and more sophisticated feature interactions (e.g., multi-variable terms) to enhance predictive performance.

References

[1] Titanic. Wikipedia. https://en.wikipedia.org/wiki/Titanic.

[2] Startupsci. Titanic data science solutions. Kaggle, 2019-04-17. https://www.kaggle.com/code/startupsci/titanic-data-science-solutions.

[3] Ai Y. Predicting Titanic survivors by using machine learning. Highlights in Science, Engineering and Technology, 2023, 34: 360–367.

[4] Ibrahim A A, Abdulaziz R O. Analysis of Titanic disaster using machine learning algorithms. Engineering Letters, 2020, 28 (4).

[5] Bisht M, Singh A, Tripathi G, Shantanu K, Gupta A, Gupta R. Analysis of machine learning algorithms for predicting Titanic disaster survival rate. 2024 4th International Conference on Sustainable Expert Systems (ICSES), 2024: 1781–1786. IEEE.

[6] Thomas T, Rajabi E. A systematic review of machine learning-based missing value imputation techniques. Data Technologies and Applications, 2021, 55 (4): 558–585.

[7] Dhal P, Azad C. A comprehensive survey on feature selection in the various fields of machine learning. Applied Intelligence, 2022, 52 (4): 4543–4581.

[8] Oh S. Feature interaction in terms of prediction performance. Applied Sciences, 2019, 9 (23): 5191.

[9] Anggodo Y P, Girsang A S. A novel modified binning and logistics regression to handle shifting in credit scoring. Computational Economics, 2024, 63 (6): 2371–2403.

[10] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. Springer, 2013.